The WWW as a Resource for Lexicography

Gregory Grefenstette

The WWW as a Resource for Lexicography

Gregory Grefenstette

Principal Research Scientist, Clairvoyance Corp., Pittsburgh, PA.

Until the appearance of the Brown Corpus with its 1 million words in the 1960s and then, on a larger scale, the British National Corpus (the BNC) with its 100 million words, the lexicographer had to rely pretty much on his or her intuition (and amassed scraps of papers) to describe how words were used. Since the task of a lexicographer was to summarize the senses and usages of a word, that person was called upon to be very well read, with a good memory, and a great sensitivity to nuance. These qualities are still and always will be needed when one must condense the description of a great variety of phenomena into a fixed amount of space.

But what if this last constraint, a fixed amount of space, disappears? One can then imagine fuller descriptions of how words are used. Taking this imaginative step, the FrameNet project has begun collecting new, fuller descriptions into a new type of lexicographical resource in which '[e] ach entry will in principle provide an exhaustive account of the semantic and syntactic combinatorial properties of one "lexical unit" (i.e., one word in one of its uses).' (Fillmore & Atkins 1998) This ambition to provide an exhaustive accounting of these properties implies access to a large number of examples of words in use. Though the Brown Corpus and the British National Corpus can provide a certain number of these, the World Wide Web (WWW) presents a vastly larger collection of examples of language use. The WWW is a new resource for lexicographers in their task of describing word patterns and their meanings. In this chapter, we look at the WWW as a corpus, and see how this will change how lexicographers model word meaning.

The Lexicographer's Task and Corpora

In the past, lexicographers worked haphazardly from their chance encounters with written sources, collecting interesting usages of words found while reading books and newspapers, and from examples drawn from their own intuition of general language use. Then in the early 1960's, there was the first, systematic attempt to create a large corpus that was representative in some way of language

use, an effort that resulted in the Brown Corpus (Francis and Kucera, 1964), 2000 segments of 500 words from a chosen variety of published sources. This new resource gave lexicographers a snapshot of actual language use at a given instant. It could be exploited by a machine to produce key-word-in-context lists to ease the lexicographer example-gathering task. The Brown Corpus had the additional advantage of being part-of-speech tagged which allowed the researcher, after some further computer developments, to look not only for words but for syntactic patterns. The appearance of this corpus gave rise to a great body of literature in computational linguistics. (See http://citeseer.nj.nec.com/context/61645/0 for a list of some articles that exploited this resource.)

And then, partly due to Sue Atkins' efforts in the 1980s, there was the development of a new lexical resource, the British National Corpus (Leech, 1992), a language sample 100 times larger than the Brown Corpus. In addition to providing more language data, the British National Corpus innovated in collecting not only 90 million words of printed text but also 10 million words of transcribed speech. The entire corpus is part-of-speech tagged, as was the Brown the Corpus, so the lexicographer can still search over both lexical and syntactic combinations. Since rare lexical and syntactic patterns will show up more often in a larger corpus, this newer corpus exposes the lexicographer to the rarer patterns of language. But such a large amount of data also shows common patterns more clearly, since the frequency of these patterns becomes much greater and more easily distinguishable from statistical blips that appear in smaller collections. This new resource likewise gave an impetus to scientific literature in lexicography and in computational linguistics. We can easily predict that the World Wide Web which appeared on the scene in 1994 will again provide a similar enhancement for all lexicographical work, since it provides a searchable corpus of language use that is much greater still than the British National Corpus.

The World Wide Web as a Corpus, Overview

In order to establish this prediction, we examine the WWW as a corpus, talking first about its size. We will see that the WWW is a few orders of magnitude bigger than the British National Corpus, and growing.

Then we address the multilingual aspect of the WWW: the fact that the WWW as a corpus contains not just English but many world languages. This mixture is both a handicap to lexicographers, since the language of text must be identified before it can be processed, but also a godsend for all work involving non-English lexicography. Non-English corpora were extremely rare and

difficult to obtain before the WWW. Diachronic studies of language on the WWW show that English, though preponderant on today's Web, is losing its pre-eminence to other languages on the WWW.

If we are to use the Web to create lexical resources for English and other languages, we need to know how clean the WWW is as a corpus of language use. One of the advantages of 90% of the British National Corpus, and of the Brown corpus, is that it is a collection of written text, often newspaper or journal text having undergone a certain amount of editing, and ensuring a certain level of correctness. Text on the WWW, on the other hand, has been created by a variety of people of different educational levels, sometimes writing in languages other than their native tongue. Below, we show, at least anecdotally, that although the Web is "dirty" with spelling errors and ungrammaticalities, correct forms are found more often than errors. The signal of correct language is much greater than the noise than these errors generate.

Further on, we talk about what the WWW allows us to do with respect to lexical combinations. It allows us to consider not just single words, but also phrasal combinations which though rarer than individual words, appear so often in this large corpus that we can apply lexicographic techniques to them. For example, we can apply KWIC, usually used for seeing how individual words are used with other words, to phrasal combinations. Benefiting from the size of the corpus that the WWW provides, we are able to consider a different dimension of what the lexicon is.

We move beyond lexicons describing individual words into very large lexicons, into the dimension of how phrases are used with other words and with other phrases. These very large lexicons can be derived for different languages, for different domains, and at different periods from a renewable resource as the WWW grows and expands.

Corpus Size: How much text is in the Web?

It is commonly admitted that "corpus-based lexicography gives a strong and necessary empirical evidence to the lexicographer's personal intuition" (Verlinde and Selva, 2001).

This belief was one of the motivations for creating the British National Corpus (http://info.ox.ac.uk/bnc/) in the early 1990's. The BNC, as mentioned above, is a large and balanced corpus, 100 million words of English drawn from thousands of sources: books, magazines, letters, medical documentation, ephemeral writings such as advertisements, as well as 10 million words of spoken English collected from volunteers over a two-week period. The BNC

provides a large, static version of English language use. The advantage for a lexicographer that such a resource provides is clear: "a large corpus reveals recurrent patterns that might not have been apparent in earlier, less comprehensive databases" (Michael Rundell, http://www.longman-elt.com/dictionaries/corpus/lrcorpus1.html).

Though the BNC is large and has the advantage over the Web of being a balanced corpus from known sources, the WWW provides a much larger corpus of English. We can get an idea of just how much larger from a simple experiment. Take some noun phrases, and see how often they appear both in the British National Corpus and on the Web. (Many web portals display how many times they have indexed words or combinations of words.) The table below gives some randomly chosen noun phrases and, in the second column, their counts in the BNC (singular and plural forms), and, in the third column, the counts for the same phrases on the Altavista web portal on a given day in late fall 2001:

PHRASE	BNC Count	ALTAVISTA Count
medical treatment	414	627522
prostate cancer	39	518393
deep breath	732	170921
acrylic paint	30	43181
perfect balance	38	35494
electromagnetic radiation	39	69286
powerful force	71	52710
concrete pipe	10	21477
upholstery fabric	6	8019
vital organ	46	28829

We see that there are about three orders of magnitude (a thousand times) more occurrences of these phrasal patterns on the Web. But this gives only an idea of the lower bound of the size of the Web. Consider that it has been estimated that Altavista visits only part of the visible Web, and that the hidden Web (Kautz et al, 1997) (i.e. the Web not indexed by web spiders, but only accessible through dialogue boxes on Web pages such as MedLine at http://www4.ncbi.nlm.nih.gov/PubMed/) is not indexed at all by these browsers and thus not counted, then the size of the corpus becomes truly staggering.

The Multilingual Corpus

In the last section, we gave a rough estimate of the lower bound of the number of words in English available on the Web. We can give a more precise estimate of the number of words that a Web portal like Altavista (www.altavista.com)

actually has indexed, not only for English but for other languages that lexicographers will work on. Here is how this estimate can be made. Start with the observation that function words, such as "the", "with", "in", etc., occur with a frequency that is relatively stable over many different types of texts. Given this stability, if we know the frequency of these function words in a corpus, then we can estimate the size of the entire corpus. As a simple example, the English word "the" appears 5,776,487 in the 90 million written text section of the in the British National Corpus (or about 7 times out every 100 words). The American Declaration of Independence contains 84 times the word "the". Knowing only this fact, we can predict that the Declaration is about 1200 (84 times 100/7) words. In fact, it contains about 1500 words. Just knowing the frequency of one word, we get the right order of magnitude.

In the last example, we used only one word "the" to predict the size of a text. If we use more words, then our prediction gets better. We can consider each word's prediction of the total size of a corpus, and then average those predictions. Here is an example using more than one word as a predictor. From a large corpus of German of a known size, we can calculate the relative frequency of the most frequent common German words. We can then formulate a query composed of these words to be submitted to Altavista. Given this query, Altavista returns a results page that shows how many times it has indexed each of these words (this information appears at the bottom of the results page). We then take the relative frequency of each word from the original known corpus, as we did above with the "the" example for English. Given the known relative frequency and the actual frequency of the word from Altavista, each word predicts the size of the entire corpus of German words. We throw out the highest and lowest predictions and average the results to get our estimate of the size of the German WWW corpus that Altavista has indexed.

Here are the frequencies that Altavista gave for some common German words in February 2000:

```
daß: 7990333; durch: 8250898; einer: 9315833; wir: 9590451;
wie: 9844516; wird: 11286438; sind: 11944284; zur: 12232738;
oder: 13566463; aus: 13678143; auch: 15504327; werden: 16375321;
sich: 17547518; nicht: 18294174; eine: 19739540; auf: 24852802;
ist: 26429327; für: 33903764; von: 39927301; und: 101250806
```

If we produce a table listing each word in the first column, the relative frequency of this word from a known German corpus in the second column, the Altavista frequency in the third, we can put in a fourth column the prediction

that one can make of the entire German Altavista corpus. Sorting this table by the final column produces a results table, part of which looks the following:

WORD	RELATIVE FREQUENCY	ALTAVISTA FREQUENCY	SINGLE WORD PREDICTION OF ENTIRE GERMAN CORPUS
...	...		...
oder	0.00561180	13,566,463	2,417,488,684
sind	0.00477555	11,944,284	2,501,132,644
auch	0.00581108	15,504,327	2,668,062,907
wird	0.00400690	11,286,438	2,816,750,605
nicht	0.00646585	18,294,174	2,829,353,294
eine	0.00691066	19,739,540	2,856,389,983
sich	0.00604594	17,547,518	2,902,363,900
ist	0.00886430	26,429,327	2,981,546,991
auf	0.00744444	24,852,802	3,338,438,082
und	0.02892370	101,250,806	3,500,617,348
...	...	...	

Average Prediction over all words except outliers: 3,068,760,356

We eliminate outliers (words which have counts that are too extreme) because Altavista does not record in its index the language a word comes from (so the counts of the string "die" include both the German and English word counts), and also because a word might be over-represented in our German training corpus, which makes their predictions too low in the Web. When we average the remaining predictions, we get a rough estimate of 3 billion words of German that could be accessed through Altavista on that day in February 2000. This technique has been tested on controlled data (Grefenstette and Nioche, 2000), in which corpora of different languages were mixed in various proportions, and gives reliable results. We estimated, in this way, the number of words that were available in 30 different Latin script languages through Altavista in March 2001. Here are those estimates:

	Word count estimate
Albanian	10,332,000
Breton	12,705,000
Welsh	14,993,000
Lithuanian	35,426,000
Latvian	39,679,000
Icelandic	53,941,000
Basque	55,340,000
Latin	55,943,000
Esperanto	57,154,000
Roumanian	86,392,000
Irish	88,283,000
Estonian	98,066,000
Slovenian	119,153,000
Croation	136,073,000

Malay	157,241,000
Turkish	187,356,000
Catalan	203,592,000
Slovakian	216,595,000
Polish	322,283,000
Finnish	326,379,000
Danish	346,945,000
Hungarian	457,522,000
Czech	520,181,000
Norwegian	609,934,000
Swedish	1,003,075,000
Dutch	1,063,012,000
Portuguese	1,333,664,000
Italian	1,845,026,000
Spanish	2,658,631,000
French	3,836,874,000
German	7,035,850,000
English	76,598,718,000

We can note that English leads the pack with 76 billion words. Still, remark that German, French, Spanish, Italian, Portuguese, Dutch and Swedish all have over a billion words accessible through Altavista. Performing the above estimations over time shows that the proportion of non-English text to English is growing over time (Grefenstette & Nioche, 2000). In October 1996 there 38 German words for every 1000 words of English indexed by Altavista. In August 1999, there were 71 German words for every 1000 English words; in March 2001, there were 92 German words for every 1000 English words.

Going back to the last table, we see that even languages such as Malay, Turkish, Slovenian and Polish present more than the one hundred million words that were collected for the British National Corpus. Some of the research that has been undertaken on the BNC that relies on its scope (though not necessarily on its being part-of-speech tagged and balanced) should be transferable to these languages. But, we repeat, it must be emphasized that these numbers are merely a lower bound on the number of words available for these languages from the Web. This is so because (a) Altavista only covers a fraction of the indexable web pages available (estimated at 15% by Lawrence and Giles (1999)), (b) Altavista might be biased to North American (mainly English language) pages by the strategy it uses to crawl the Web, and (c) Altavista only indexes pages that can be directly called by a URL, and does not index all the text can be found in databases that are accessible through dialog windows on Web pages.

It is clear from these size estimates that, given the proper tools, the "recurrent patterns" that it is the lexicographer's role to describe can be found for many languages using the Web as a corpus. The maximalist list of tools, seen below, that are needed to automatically retrieve and package these patterns are web

crawlers, language identifiers, domain and genre classifiers, and morphological analyzers, parts of speech taggers and shallow parsers. Some of these tools are language independent, and some are not. All the tools exist for English, and many are being developed for other languages. The language dependent, linguist tools can be approximated, of course (Grefenstette, 1998). But whether the tools exist now or not, it is evident that they will have to be constructed since English is not becoming the only language on the WWW. These other languages are not going away. The lexicographical work that has been performed in English, thanks to resources such the BNC, can and must be performed on non-English languages as well. The multilingual Web provides the raw material.

The Dirty Web

So, the WWW is big, but it is obviously not as clean as a corpus of newspaper texts: people using the Internet may be writing their texts in a non-native language; they may be using incorrect speech; and they will be making unedited grammatical and spelling errors. Of course, the errors that people make interest a certain branch of psychology (Gass, 1979) but do these errors prevent using the Web as a lexicographic resource? Can the Web as a corpus be useful for a lexicographer wishing to produce a "correct" version of how language is used? We can make some anecdotal observations in response.

Every language contains forms that are considered erroneous by the majority of educated speakers. For example, there is a common error made by Spanish users labeled as a dequeismo, meaning to place a spurious "de" between a verb and its following relative clause. Below, we give some examples of this error, and the correct forms. Each pattern is followed by the count of the sequence (the number of pages it is found on) from the pages that have been indexed by AllTheWeb. Even without understanding Spanish, it is easy to guess which is the correct form and which is erroneous.

```
"pienso de que"   171 times
"pienso que"    83966 times
"piensas de que"   89 times
"piensas que"   11485 times
"piense de que"     9 times
"piense que"    12867 times
"pensar de que"   716 times
"pensar que"   188508 times
```

Source:www.alltheweb.com (June 2001)

One can repeat this small experiment with other languages. Some common grammatical errors from Dutch involve choice of prepositions. Here are some

examples of the correct and incorrect patterns with their Web counts. Again, it is easy to see which is the correct form.

"hun hebben het"	10 times
"ze hebben het"	2459 times
"groter als"	1079 times
"groter dan"	20421 times
"betreffende hen"	12 times
"betreffende hun"	329 times
"behalve hen"	12 times
"behalve hun"	310 times

Source: www.alltheweb.com (June 2001)

The same phenomena can be found by looking at incorrectly and correctly spelled words. Some examples from English: in June 2001, there were 1692 "I beleave", 41617 "I beleive" but 3,800,810 correct forms "I believe" to be found on English language web pages.

In all these anecdotal cases, one could pick the correct form without knowing the languages, simply using the counts from the Web. The correct form is always orders of magnitude more frequent than the erroneous form. The WWW can be a source for modeling how language is correctly used if thresholds are applied. The Web is dirty but the signal (correct forms and correct usage) is so strong noise can easily be ignored.

Tools for Extracting Recurrent Patterns from the Web

In order to exploit the Web for lexicography, a certain number of existing tools have to be strung together. Pages from the web can be automatically collected using Web Crawlers (Heydon and Najork, 1999). Crawlers, sometimes called spiders, are programs that fetch the pages indicated by a URL (Universal Resource Locator), e.g., http://info.ox.ac.uk/bnc/. The crawlers fetch all the pages in a list of URLs, analyzing each page fetched and adding to that list any new URLs found on an accessed page. When the list is empty, it can be refilled by generating random internet addresses (e.g., http://123.45.67) some of which will correspond to real pages, thus restarting the process. Instead of using a web crawler to collect text, one can also parasite commercial portals such as www.google.com, www.alltheweb.com, www.yahoo.com, www.altavista.com, etc. by formulating queries involving the words to be studied and collecting all the URLs returned. Most portals return a maximum of 1000 URLs for a given query. The UNIX command "wget" can be used to build a Web Crawler.

Once Web pages associated with URLs are fetched, these pages must be converted to ordinary text using options such as "Save as Text" on web browsers like Netscape and Internet Explorer, or by using programs such as "dehtml" (see http://www.moria.de/~michael/dehtml/). SGML marks such as " " (non-breaking space) which code non-ASCII characters such as accented characters, mathematical symbols, etc., must be converted into single character code. See a list of these conversions at http://www.iro.umontreal.ca/contrib/recode/charsets/rmail/html.

Given a textual representation, a language has to be associated with each fetched and converted page. Though there are XML codes (http://www.w3.org/International/O-HTML-tags.html) that allow the creator of a page to specify the language of the page, these codes are rarely used in the anarchic world of the WWW. Language codes can be automatically associated to a page using a language identifier (Grefenstette, 1995) that models the common character sequences of the languages found on the Web. The models are derivable from a small training corpus (e.g., 1 Mbyte of text in each language to be recognized) by calculating the relative frequency of all the sequences of two, three, or four characters found in that language. Keeping the most frequent 1000 sequences with their frequencies is sufficient to distinguish languages. These lists will show, for example, that the word terminal sequence "ing" is more common in English than in French, while the sequence "que" is more common in French. Given a new page found on the Web, one extracts the character sequences from this new unknown language page and matches them with the sequences from the list of known languages, picking the best match as being the language of the new page. A small number of Web pages are truly multilingual, but the same language identification procedure can be applied on a paragraph level, or even on a sentence level, though one needs about twenty to thirty words to identify a passage with high accuracy.

Once a page has been fetched and its language identified, the domain of the page (e.g., finance, medicine, sports, etc.) can optionally be identified before the lexicographic process begins. Whereas language identifiers use sequences of characters, a domain identifier (Nigam et al, 2000) uses the presence of words that are characteristic of the domain. To build up a domain identifier, one needs a large collection of text known to belong to that domain and a large collection of text known not to belong to the domain. One builds from these collections lists of words that appear more often in the domain than not in the domain. These words and their frequencies become the model of the domain. Given a new text, one extracts all the words (lemmatizing, stemming or merely

truncating the words after five or six characters in order to improve recall) and matches these words against the domain profiles, choosing the profile that matches most closely. By fetching and classifying a large number of pages, one can collect as large a corpus as desired. This corpus can then be used to extract the recurrent patterns proper to a language or to a given domain within a language.

The linguistic tools necessary for optimal recognition and extraction of lexical and syntactic patterns do not exist yet for all languages. In (Grefenstette, 1998) we show how one can approximate parsing, progressing from simpler to more elaborate tools and methods. One can approximate parsing in a graduated fashion using simple tokenization, using a small list of function words, using positional information, using part-of-speech information, using sequences of part of speech tags, up to using shallow parsing. With each new linguistic resource added, the patterns present in text become clearer to isolate automatically, and clearer to see for a human. For example, suppose that the lexicographer is working on the word "check." Using two simple tools, a program that looks in a three word window and one that identifies words from a list of preposition, one finds in the BNC following the word "check" the following prepositions:

Frequency	preposition
1050	on
652	for
619	of
582	with
507	to

With these two linguistically simple tools, one can begin to see the important recurrent patterns involving "check". If one further applies a part-of-speech tagging program, one can find within three word window after "check", within three words after a window with "check...on", and within three words after a window containing "check..for" the following nouns in the BNC:

Nouns after check...		***after check .. on ..***		***after check .. for ...***	
97	watch	28	progress	21	sign
51	time	12	movement	14	error
44	number	11	thing	12	accuracy
43	progress	11	number	11	leak
37	record	9	file	11	damage
35	detail	8	use	10	check
33	thing	8	time	8	consistency

33	list	8	quality	7	time
32	accuracy	8	people	7	possible
31	fact	8	level	7	level
30	check	8	activity	7	correct
29	level	7	make	5	square
29	information	7	car	5	pulse
27	name	6	record	5	free
27	date	6	performance	5	flight
27	car	6	material	5	fingerprint
25	hotel	6	health	5	detail
23	work	6	gas	5	crack
23	spelling	6	child	4	wear
22	file	6	calculator	4	virus

We see that one "checks" watches, the time, progress; details, lists, records, files; names, information, the car, levels, spelling, etc. One "checks on" the progress; one "checks for" signs and "checks for" fingerprints. The recurring patterns become clearer than they would be if one took all strings appearing in the same window. They become clearer because we are approximating parsing (which would tell use exactly what the arguments of "check" are), but this approximation uses simpler tools in lieu of a full-blown parser which might not exist, or be accessible, for the language in question. As one continues to add more evolved linguistic tools, the patterns become more refined. Using a shallow parser to extract patterns (Grefenstette, 1996), we see an improvement in the recognition of the patterns as shown in the table below which shows the first few arguments of "check...with" compared to simply finding words appearing after "check...with":

Parsed		3 words after	
check ... with ...		check ... with ...	
20	office	19	local
8	doctor	19	doctor
7	authority	13	office
5	agent	12	level
4	number	9	spirit
4	manager	7	bank
4	detector	6	manager
. . .		. . .	

As one adds in more linguistic knowledge in the form of more evolved linguistic processing, rarer phenomena rise to visibility. These shallow, robust parsing tools, used in the last example, are becoming more accessible in a number of languages (Ait-Mokhtar and Chanod, 1997). It is possible for

lexicographers to incorporate these tools in their work, producing KWIC indexes over shallowly parsed results as shown below. Here are KWIC lines associated not just with "check" but with the lexical syntactic pattern of "checking with" either a "bank" or "editor" or "boss." It is possible to produce these resources for lexicographers because (a) we are using a large corpus (though the BNC is a little stretched for this task) and (b) we integrate a shallow parser into the automatic pre-treatment of the text.

shops and other establishments abroad	check with your bank for details .
So it is essential to	check with your travel agent or bank on
No, cheque - but it 's good - Josh	checked with the bank, called Hnatiuk
this was a good idea but would have to	check with his immediate boss
Universe editor Anne Noels	checked with bosses about the ban in
It is sensible, however, to	check with editors of really specialist
On being told by the manager to	check with the bank, he pretended to
Then, posing as a relative, she	checked with the editor .
She	checked with all the contributing edit...
with drivers stranded in France	checking in with their boss .

If this same treatment were run over the Web, we would find more than 7000 instances of "checking with bank" and thousands of other instances of "checking with editors" or "with bosses." We would be able to perform comprehensive lexicographic work on the more precise pattern.

Word Sketches and Very Large Lexicons

Given a large corpus, such as the WWW, it is possible to extract large descriptions of how lexical patterns are used. An early attempt at automatically creating a large lexicon involving lexical and syntactic patterns was undertaken by the DECIDE project (Grefenstette et al, 1996). This project (1994-96) used shallow parsing for English, French and German corpora to extract recurrent lexical patterns which were embedded into lexicons whose entries corresponded to phrasal patterns. The results can be seen at
http://engdep1.philo.ulg.ac.be/decide/lexicon/.

Here is part of the lexicon for "pay a compliment." The lexicon entry contains the frequency of the pattern in the reference corpus (here 23 times), as well as other information about the pattern, such as what prepositions followed the pattern (here "to" 7 times and also "on" and "in"), the determiners preceding "compliment" with their frequencies, the voice (active, passive, past participle) of the verb, typical excerpts containing the pattern, and syntactic patterns found when "compliment" is used as a verb.

```
<collocate> compliment ; pay;  DOBJ (23) </collocate>
...
<subcat>
 (7)  pay compliment to
 (1)  pay compliment on
 (1)  pay compliment in
</subcat>
<noundet>
the compliment (9)
NONE compliment (7)
a compliment (3)
her compliment (2)
my compliment (1)
his compliment (1)
</noundet>
<voice>
VOICE=ACTIVE (11)
VOICE=INFINITIVE (7)
VOICE=PPART (6)
VOICE=PASSIVE (5)
</voice>
<typical>
   2 ... pay compliment...
   1 ... the usual compliment pay to my unimprovable English...
   1 ... the ultimate compliment be pay...
   1 ... the most kindly and satisfying compliment to the
Cornish be pay...
   1 ... the many compliment can be pay...
</typical>
   <verbequiv>
  compliment
   <vsubcat>
   (124) compliment NP
   (70) compliment ... on
   (20) compliment NP on
   (7) compliment ... by
   (6) compliment ... in
   (5) compliment ... of
   (5) compliment ... for
   (4) compliment NP of
   (4) compliment ... to
   (3) compliment NP in
  </vsubcat>
 </verbequiv>
```

The purpose of this lexicon was to automatically construct a large image of how some multiword lexical patterns were used. More recently a larger-scale project is being run to produce more elaborate Word Sketches (Kilgarriff & Tugwell, 2001) at the ITRI by Adam Kilgarriff (see also in this volume) and David Tugwell with direction and advice from Roger Evans and Sue Atkins.

Such automated techniques will find their full value when run against the WWW, because, as we have seen, the size of the Web will make common recurrent patterns easier to identify using simple linguistic techniques, and it will bring up to visibility rarer patterns not present in smaller corpora. We will be able to produce very large lexicons of word sketches for different languages, for different domains, and for different genres. For example, here below we have the outline of a hypothetical entry for the pattern "presidential election." We imagine that it has been constructed from text that has been classified as belonging to the domain of politics. The entry does not have a headword but a head dependency relation (ADJ means an adjective modifying a noun). The entry includes the relative frequency of this relation in the corpus. This frequency would be useful for applications such as speech recognition and OCR in which it could be used to decide between two alternative readings of a text. After the frequency there is a list of other dependency relations that are variants of the head relation (here we have DOBJ, "electing a president" as direct object, the passive subject form (SUBJPASS), the prepositional noun phrase form (NNPREP), the past participle form (NPDOBJ), etc.). Then, the entry contains a context section with the most frequently co-occurring words and relations within some window after the appearance of the relation. Follows a section showing the entities (people, places and things) that are found in these windows. The last section (though one can conceive of more sections) would point to other dependency relations involving the lexical items in the entry, glossed (Grefenstette, 1997) here as other "presidential things," and other types of "elections."

```
LEXICON:    Politics
ENTRY:      ADJ(presidential election)
FREQUENCY:  2,486/100,000,000
VARIANTS:   DOBJ(elect president)
            SUBJPASS (president was elected)
            NNPREP(election of president)
            NPDOBJ(elected president)
CONTEXT:    50 words (frequency > 5) before/after other entries
            found more than once in window,
            e.g. NN(acceptance speech)
```

```
ENTITIES:  other recognized people, places, things
NETWORK:   pointers to lexical class members
           ADJ(presidential, * ) presidential things
           ADJ( *, election) types of elections
```

Conclusion

The preceding discussion about the size of the WWW, its multilinguality, its usefulness for extracting clean lexical patterns, and the extraction of lexical patterns from corpora show that we will be able to automatically build large, organized collections of word usages for specific languages and for specific domains. What implications does this have for the lexicographer? Of course, the lexicographer's task will still be to produce a generalization and a succinct description of the nuances between word and phrasal choice, now from a much more complex representation of attested word use. Rather than working from the display of KWIC indexes, intuiting the common thread between similar word uses, the future lexicographer will be involved in creating these abstractions with a computer. The future lexicographer, mastering these text manipulation and text processing tools from a programming point of view will creating new types of representations for both human and computer consumption. The future lexicographer will be not only master his or her language, but will also master the computer. Lexicography will be performed by computational lexicographers rather than by language artisans.

References

Ait-Mokhtar, S. and Chanod, J. 1997. Incremental Finite-State Parsing. In Proceedings of ANLP-97, Washington, DC, pp. 72-79

Fillmore, C. and Atkins, B.T.S. 1998. "FrameNet and Lexicographic Relevance", First International Conference on Language Resources & Evaluation: Proceedings, pp. 417-420.

Francis, W. N. and Kucera, H. 1964. "Brown corpus manual: manual of information to accompany a standard corpus of present-day edited American English, for use with digital computers", Brown University, Providence, Rhode Island.

Gass, S. 1979. "Language transfer and universal grammatical relations". Language Learning, 29(2):327—344

Grefenstette G. 1995. "Comparing two language identification schemes". Proceedings of the 3rd International Conference on the Statistical Analysis of Textual Data (JADT'95), Rome, Italy, pp. 263-268.

Grefenstette, G. 1996. "Light parsing as finite-state filtering". In Kornai A., editor, Proceedings of the ECAI 96 Workshop on Extended Finite State Models of Language, pages 20--25.

Grefenstette, G. Heid, U. Schulze, B.M. Fontenelle, T. and Gerardy, C. 1996. "The DECIDE Project: Multilingual Collocation Extraction", EURALEX'96 Proceedings, University of Göteborg, pp.293-107

Grefenstette, G. 1997. "Short Query Linguistic Expansion Techniques: Palliating One-Word Queries by Providing Intermediate Structure to Text". In Pazienza, M. T., editor, Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, Springer- Verlag, pp. 97-114.

Grefenstette, G. 1998. "The Future of Linguistics and Lexicographers: Will there be Lexicographers in the year 3000?" Euralex '98 Proceedings, Liege, Belgium, vol. 1, pp. 25-41.

Grefenstette, G. and Nioche, J. 2000. "Estimation of English and non-English language use on the WWW". Proc. RIAO 2000, Content-Based Multimedia Information Access, pages 237-- 246.

Heydon; A, and Najork, M. A. 1999. A scalable, extensible web crawler. World Wide Web, 2(4):219-229

Kautz, H., Selman, B., Shah, M. 1997. "The Hidden Web", The AI Magazine, Vol.18, No.2, pp27-36.

Kilgarriff, A. and Tugwell, D. 2001. "WORD SKETCH: Extraction and Display of Significant Collocatioins for Lexicography". ACL workshop "COLLOCATION: Computational Extraction, Analysis and Exploitation". Toulouse, to appear.

Lawrence, S. and Giles, C. L. 1999. "Accessibility of information on the web", Nature 400, 107–109.

Leech, G. 1992. 100 million words of English: the British National Corpus. Language Research 28(1), pp. 1--13.

Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. M. 2000. "Text classification from labeled and unlabeled documents using EM". Machine Learning, 39, 103--134.

Verlinde S., Selva T. 2001. "Corpus-based vs. intuition-based lexicography: defining a word list for a French learners' dictionary," Proceedings of the Corpus Linguistics 2001 conference, Lancaster University (UK), pp. 594-598.